IAT 2 - Natural Language Processing (NLP) - Question Paper

USN | | | | | | | | | | |

Internal Assessment Test 2 – May 2025

Sub:	Natural Language Processing					Sub Code:	BAD613B	Branch:	AI&DS
Date:	23/05/2025	Duration:	90 mins	Max Marks:	50	Sem / Sec:	VI		OBE

Answer any FIVE FULL Questions

		MARKS	CO	RBT
1a	Discuss the applications of Text Classification	[5]	CO2	L2
1b	Explain multinomial Naive Bayes Classifier and its mathematical principles	[5]	CO2	L3
2a	**Training Data**: Just plain boring (-); Entirely predictable and lacks energy (-); No surprises and very few laughs (-); Very powerful (+); The most fun film of the summer (+). **Test Data**: Predictable with no fun (?). Classify the test message as "+" or "-" using Naive Bayes.	[5]	CO2	L2
2b	Demonstrate the application of the Naïve Bayes algorithm as a language model and discuss its evaluation metrics.	[5]	CO2	L3
3a	Write a short notes on Information Retrieval (IR) models	[5]	CO3	L2
3b	Explain Boolean model with example	[5]	CO3	L3
4a	What are lexical resources, and how are WordNet, FrameNet, stemmers, POS taggers, and corpora used?	[5]	CO3	L2
4b	Summarize about Major issues in Information Retrieval.	[5]	CO4	L3
5a	Explain machine translation using the encoder-decoder architecture.	[5]	CO4	L2
5b	Discuss language divergences and linguistic typology.	[5]	CO4	L3
6a	What is machine translation evaluation? Explain its importance.	[5]	CO4	L2
6b	Describe the issues of bias and ethics in the context of machine translation.	[5]	CO4	L3

NLP Answer Key – IAT 2

Sub:	Natural Language Processing			Sub Code:	BAD613B	Branch:	AI&DS
Date :	24/05/2025	Duration:	90 mins	Max Marks:	50	Sem	VI

1a **Discuss the applications of Text Classification**

1. **Spam Detection**: Classifies emails as *spam* or *not spam* based on content and keywords.
2. **Sentiment Analysis**: Determines the *emotional tone* (positive, negative, neutral) in reviews, social media posts, etc.
3. **News Categorization**: Automatically classifies news articles into categories like *sports*, *politics*, *technology*, etc.
4. **Language Detection**: Identifies the *language* of a given text (e.g., English, French, Tamil).
5. **Topic Labeling in Customer Support**: Tags customer queries or tickets by topic (e.g., *billing*, *technical issue*, *feedback*) for quicker resolution.

1b **Explain multinomial Naive Bayes Classifier and its mathematical principles**

1. **Naïve Bayes** is a classification algorithm based on **Bayes' Theorem**, used to predict the class of given data.
2. It assumes that all input features are **independent** of each other — this is the "naïve" assumption.
3. **Bayes' Theorem**:

$$P(C|X) = \frac{P(X|C).P(C)}{P(X)}$$

Where $P(C|X)$ is the probability of class C given data X.

4. The classifier selects the class with the **highest probability** using:

$$\widehat{C} = \underset{C_k}{arg\,max} P(C_k) \prod_{i=1}^{n} P(x_i|C_k)$$

5. It is simple, fast, and works well for **text classification** problems like spam detection and sentiment analysis.

2a

Training Data		Test Data	
Message	**Label**	**Message**	**Label**
Just plain boring	-	Predictable with no fun	?
Entirely predictable and lacks energy	-		
No surprises and very few laughs	-		
Very powerful	+		
The most fun film of the summer	+		

Classify the test message as "+" or "-" using Naive Bayes.

The prior $P(c)$ for the two classes is computed via Eq. 4.11 as $\frac{N_c}{N_{doc}}$:

$$P(-) = \frac{3}{5} \qquad P(+) = \frac{2}{5}$$

The word *with* doesn't occur in the training set, so we drop it completely (as mentioned above, we don't use unknown word models for naive Bayes). The likelihoods from the training set for the remaining three words "predictable", "no", and "fun", are as follows, from Eq. 4.14 (computing the probabilities for the remai of the words in the training set is left as an exercise for the reader):

$$P(\text{"predictable"}|-) = \frac{1+1}{14+20} \qquad P(\text{"predictable"}|+) = \frac{0+1}{9+20}$$

$$P(\text{"no"}|-) = \frac{1+1}{14+20} \qquad P(\text{"no"}|+) = \frac{0+1}{9+20}$$

$$P(\text{"fun"}|-) = \frac{0+1}{14+20} \qquad P(\text{"fun"}|+) = \frac{1+1}{9+20}$$

For the test sentence S = "predictable with no fun", after removing the word 'w the chosen class, via Eq. 4.9, is therefore computed as follows:

$$P(-)P(S|-) = \frac{3}{5} \times \frac{2 \times 2 \times 1}{34^3} = 6.1 \times 10^{-5}$$

$$P(+)P(S|+) = \frac{2}{5} \times \frac{1 \times 1 \times 2}{29^3} = 3.2 \times 10^{-5}$$

The model thus predicts the class *negative* for the test sentence.

2b

Demonstrate the application of the Naïve Bayes algorithm as a language model and discuss its evaluation metrics.

1. **Language Modeling**: Naïve Bayes can be used to model the probability of words in a language and classify text into specific languages based on word occurrence.
2. **Text Classification**: It is widely used for tasks like spam detection, sentiment analysis, and language identification using word frequency and class probability.
3. **Training Process**: The algorithm calculates prior probabilities P(C)P(C)P(C) for each language and likelihood P(wi ∣ C)P(w_i | C)P(wi ∣ C) for each word wiw_iwi given a class (language).

	4. **Evaluation Metrics**: Common metrics include: **Accuracy**: Overall correctness of the model. **Precision**: Correct positive predictions out of total predicted positives. **Recall**: Correct positive predictions out of actual positives. **F1-Score**: Harmonic mean of precision and recall. 5. **Advantage**: Naïve Bayes is simple, fast, and effective for large-scale language-based classification problems with high-dimensional data.
3a	**Write a short notes on Information Retrieval (IR) models** 1. **Boolean Model (Classical):** Uses logical operators (AND, OR, NOT) for exact match retrieval. Simple but does not rank documents. 2. **Vector Space Model (Classical):** Represents documents and queries as vectors and uses **cosine similarity** to rank based on relevance. 3. **Probabilistic Model / BM25 (Classical):** Estimates the probability of document relevance to a query. BM25 is a popular ranking function based on this model. 4. **Fuzzy and Extended Boolean Models (Non-Classical):** Allow partial matches using weighted terms or fuzzy logic, improving flexibility over standard Boolean retrieval. 5. **Language Models (Non-Classical):** Treat each document as a probability distribution and rank by the likelihood that it generates the query. 6. **Latent Semantic Indexing (LSI) (Alternative):** Uses mathematical techniques (like SVD) to identify hidden relationships between terms and documents. 7. **Neural IR Models (Alternative):** Use deep learning (e.g., BERT) for semantic matching and contextual understanding, offering state-of-the-art performance.

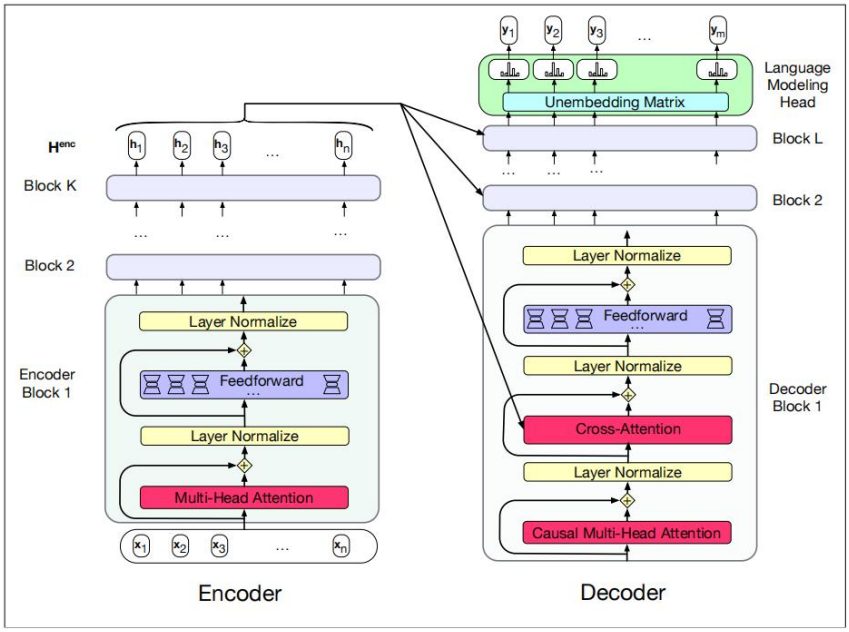
3b	**Explain Boolean model with example** 1. **Definition**: The Boolean model is a classical Information Retrieval model that represents documents and queries as sets of keywords. It uses **Boolean logic operators** like **AND, OR, NOT** to retrieve documents that match the query exactly. 2. **Working Principle**:Each document is either **relevant (1)** or **not relevant (0)** — no ranking is involved. Queries are formed using Boolean expressions to combine terms. 3. **Operators**: **AND**: Retrieves documents containing *all* specified terms. **OR**: Retrieves documents containing *any* of the terms. **NOT**: Excludes documents containing a specific term. 4. **Example**: Suppose we have a document set: D1: "apple banana mango" D2: "banana orange" D3: "apple mango" Query: `apple AND mango` Result: D1 and D3 (both contain "apple" and "mango") 5. **Advantages and Limitations**: **Pros**: Simple to understand and implement. **Cons**: No ranking of documents; can't handle partial relevance or term weighting.
4a	**What are lexical resources, and how are WordNet, FrameNet, stemmers, POS taggers, and corpora used?** 1. Lexical Resources: These are databases or tools that provide information about words, their meanings, relationships, usage, and grammatical properties. They are essential in Natural Language Processing (NLP) tasks like parsing, tagging, and semantic analysis. 2. WordNet: A lexical database of English where words are grouped into sets of synonyms (synsets) with definitions and semantic relations (e.g., synonyms, antonyms, hypernyms). Used in word sense disambiguation and semantic similarity tasks. 3. FrameNet: A resource based on frame semantics. It documents how words evoke specific

	situations (frames) and the roles (participants) in them. Used in semantic role labeling and deep NLP understanding. 4. Stemmers: Tools that reduce words to their root form by removing suffixes (e.g., "running" → "run"). Common stemmers: Porter, Snowball. Used in information retrieval and text normalization. 5. POS Taggers (Part-of-Speech Taggers): Tools that label words with their grammatical category (noun, verb, adjective, etc.). Useful in syntactic parsing, chunking, and named entity recognition. 6. Corpora: Large and structured collections of text (e.g., Brown Corpus, Penn Treebank) used to train, test, and evaluate NLP models.
4b	**Summarize about Major issues in Information Retrieval.** ✓ **Query Understanding**: Users often submit short or ambiguous queries. Interpreting the actual intent behind a query is a major challenge in IR. ✓ **Relevance Determination**: Determining what makes a document *relevant* to a query varies between users. IR systems struggle with modeling subjective relevance accurately. ✓ **Ranking of Results**: It is difficult to design effective ranking algorithms that consistently place the most relevant documents at the top. ✓ **Handling Synonyms and Polysemy**: ✓ **Synonyms**: Different words with the same meaning (e.g., "car" and "automobile"). ✓ **Polysemy**: Same word with different meanings (e.g., "bank" – riverbank or financial institution). Both cause problems in matching queries to documents. ✓ **Scalability and Performance**: With massive volumes of data on the web, IR systems must be efficient, fast, and scalable for real-time performance.
5a	**Explain machine translation using the encoder-decoder architecture.** Machine Translation is the automatic conversion of text from one language to another. The **encoder-decoder architecture** is a popular deep learning approach used for this task. 1. **Encoder**: The encoder processes the input sentence (source language) word by word using models like RNN, LSTM, or GRU. It converts the entire sentence into a fixed-length context vector (a summary of the input). 2. **Context Vector**: This vector captures the meaning of the whole sentence and is passed to the decoder. 3. **Decoder**: The decoder takes the context vector and generates the output sentence (target language) one word at a time. It predicts the next word based on the context and the previously generated words. 4. **Training**: The model is trained using parallel corpora (pairs of sentences in

source and target languages) to minimize translation errors.

5. **Limitations and Improvements**: A basic encoder-decoder struggles with long sentences. To improve this, **attention mechanisms** are used, allowing the decoder to focus on relevant parts of the input sentence at each step.

This architecture forms the foundation for many modern neural machine translation systems.



The transformer block for the encoder and the decoder.

The transformer-based encoder-decoder architecture for machine translation consists of two main parts: the **encoder**, which processes the source sentence into contextual embeddings (Henc), and the **decoder**, which generates the target sentence one word at a time. The decoder includes an extra **cross-attention layer** that allows it to attend to all encoder outputs, using them as keys and values while generating each word. This attention helps align source and target tokens effectively. Training is done using **teacher forcing** and **cross-entropy loss**, where the model is trained to predict the next word using the correct previous target token. Decoding often uses **beam search** for better translation quality.

5b **<u>Discuss language divergences and linguistic typology.</u>**

- **Language Divergences** refer to the structural and semantic differences between languages that can make machine translation challenging. These include:
  - ✓ **Lexical divergence**: One language may have a single word, while another may need a phrase (e.g., "uncle" in English vs. separate words for maternal/paternal uncle in Hindi).
  - ✓ **Syntactic divergence**: Differences in word order, like Subject-Verb-

	Object (SVO) in English vs. Subject-Object-Verb (SOV) in Japanese. ✓ **Morphological divergence**: Some languages use rich inflection (e.g., Turkish), while others use word order or prepositions. ➢ **Linguistic Typology** is the classification of languages based on their common structural features. It helps in understanding: ✓ Word order types (SVO, SOV, etc.) ✓ Morphological types (isolating, agglutinative, fusional) ✓ Syntactic and grammatical patterns
6a	**<u>What is machine translation evaluation? Explain its importance.</u>** **Machine Translation (MT) Evaluation** is the process of assessing the quality of automatically translated text. It determines how well the translated output preserves the meaning, fluency, and grammatical correctness of the source text. There are two main types: ✓ **Automatic Evaluation**: Uses metrics like BLEU, METEOR, and ROUGE to compare machine output with human reference translations. ✓ **Human Evaluation**: Involves human judges rating translations based on adequacy (meaning preservation) and fluency (naturalness of language). **Importance**: ✓ Ensures the **accuracy** and **reliability** of translations. ✓ Helps in **comparing different MT systems**. ✓ Guides **model improvement** by identifying errors. Essential for real-world applications like healthcare, legal, and international communication where precision is critical.
6b	**<u>Describe the issues of bias and ethics in the context of machine translation.</u>** Machine Translation (MT) systems can unintentionally reflect and amplify **biases** present in their training data. These biases may be related to **gender**, **culture**, or **social norms**. For example, translating gender-neutral words from one language to another may result in gender-stereotyped outputs (e.g., translating a neutral profession in a source language to "he is a doctor" or "she is a nurse"). **Ethical issues** include: ✓ **Fairness**: MT should treat all groups equally and avoid reinforcing stereotypes. ✓ **Transparency**: Users should be informed when translations are

	machine-generated. ✓ **Accountability**: Developers must ensure responsible use, especially in sensitive domains like healthcare or law. ✓ **Data Privacy**: MT systems should not leak or misuse user data. Addressing bias and ethics is crucial to ensure that MT tools are **inclusive, fair, and trustworthy** in global communication.